# Real Estate Market Insights: An Exploratory Analysis of Zameen.com Listings in Pakistan

EDA Portfolio Project
By Khalida Salahuddin
DA6 atomcamp

## Section 1: Project Objective

To extract actionable insights from property listings on Zameen.com such as pricing trends, neighbourhood comparisons, and listing quality that can help real estate investors make informed decisions.

### Problem Statement

Define the business question: What drives property prices in Pakistan?

Property prices in Pakistan are known to be influenced by the following factors:

1. Land size, land location, land use (residential and commercial)

2. Land administration

3. Construction costs which include labor and material

4. Building specifications and amenities

It's important to use real estate data to analyse variables that affect the pricing for informed decision making by the stakeholders i.e. property buyers, sellers, developers and government regulatory bodies and policy makers.

I will analyse the effect of variables on property prices with a scraped real estate dataset from Zameen.com.

## Section 2: Data Cleaning and Processing

### Initial Steps:

1. Conducted this analysis on Google Colab
2. Imported Pandas and Numpy Library
3. Read the dataset from google drive

**Understanding the data**

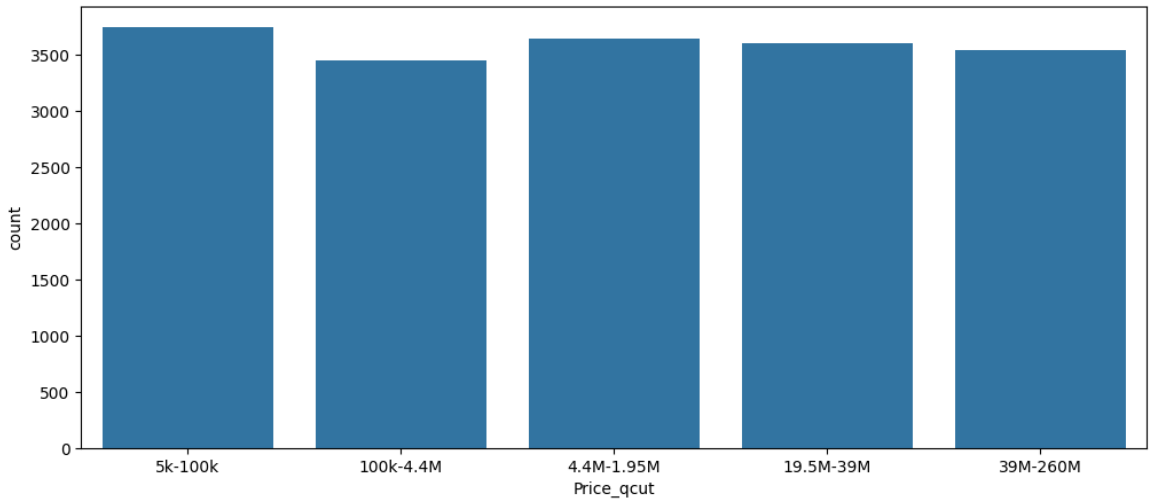
1. The dataset shows real estate listings for rent and sale in 56 unique cities of Pakistan
2. The dataset consists of 18,255 rows and 59 columns before data cleaning and 17,978 rows and 22 columns after data cleaning
3. There were no duplicates in the data
4. From 56 columns, 30 completely null columns and 1 column with null values greater than the 50% threshold values were deleted
5. For null rows, a mode value was added in 5 numeric columns and a placeholder text (not available) was added for an object datatype column
6. Wrong datatypes of 7 columns were converted into correct datatypes. Built in year only had year in it so in the absence of a complete datetime format, it was giving error for conversion to data. Hence, Built in year was converted into an integer
7. Built in year column had 537 mistype errors. Some values with single possibilities were corrected based on the closest year. While the others with multiple possibilities were replaced with mode of year
8. The price column had values in the format "PKR\n4.75 Crore". It was corrected by removing PKRn/. The price was then converted into a complete format by replacing Arab, Crore, Lacs and Thousand with zeros ('0). The datatype was changed to integer
9. The column area had property sizes in different units e.g. Sq Yd, Marla and Kanal. It was converted into a Sq Ft unit for coherence
10. One listing had an area of 45,000 kanal in Lahore which appears to be a mistype and the price is the rental value of Rs. 200,000. So the area was replaced with the area of a rental home with the same number of bedrooms and bathrooms in the same locality. It matched 1 kanal rental homes at 200,000 rent with 6 bedrooms and 5-6 bathrooms in Lahore.
11. City names were standardized using the FuzzyWuzzy Library
12. IQR and Z-score values were calculated to find outliers in prices. 261 rows with outlier prices were deleted as they had highly skewed values that would negatively affect our analysis. In real estate, sellers and property dealers are in the practice of deliberately stating inflated prices to increase the resale and rental values of their properties. Currently, market rates are heavily reliant on speculation. There is no defined range. In order, to impute values, it is important to have a defined rate for a given property which

is a lengthy process. In this project, we are only focusing on cleaning and analysis based on reasonable assumptions, so I deleted the outliers

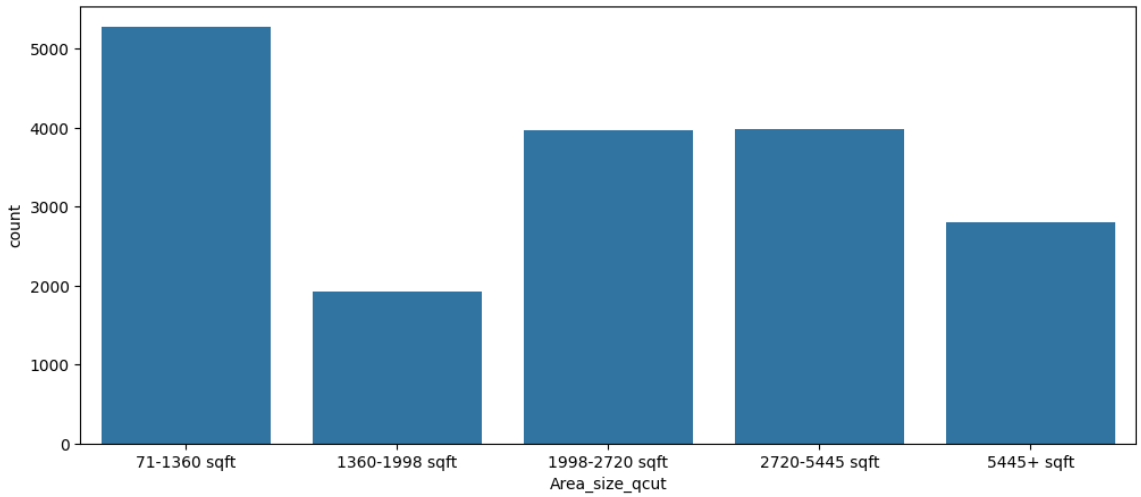
13. As a part of Feature Engineering, two new columns were created:
    1. Price/Sqft of property. It gives us a true picture of relevant property rates per unit in the market. It also shows over-inflated properties
    2. Age of property. It is helpful in assessing if the price is reasonable or inflated. Older properties were cheaper to build than today, hence the price should be lower than a new property in the same locality and size
14. Bins were created for price and area_sqft column to categorize the large data for analysis

## Section 3: Data Analysis

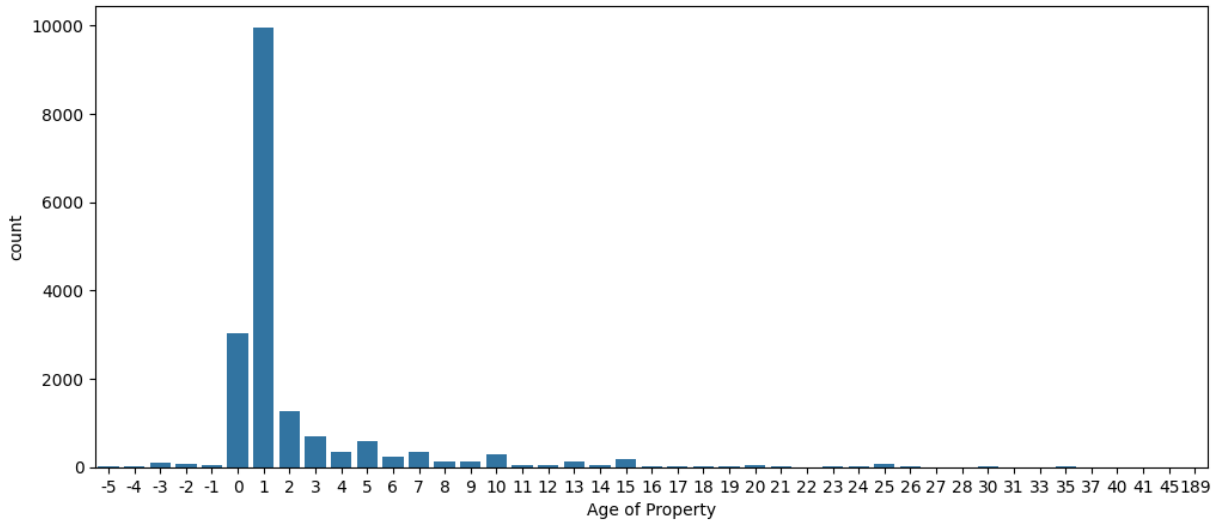
### Barplot of Count of Listing by Price Range



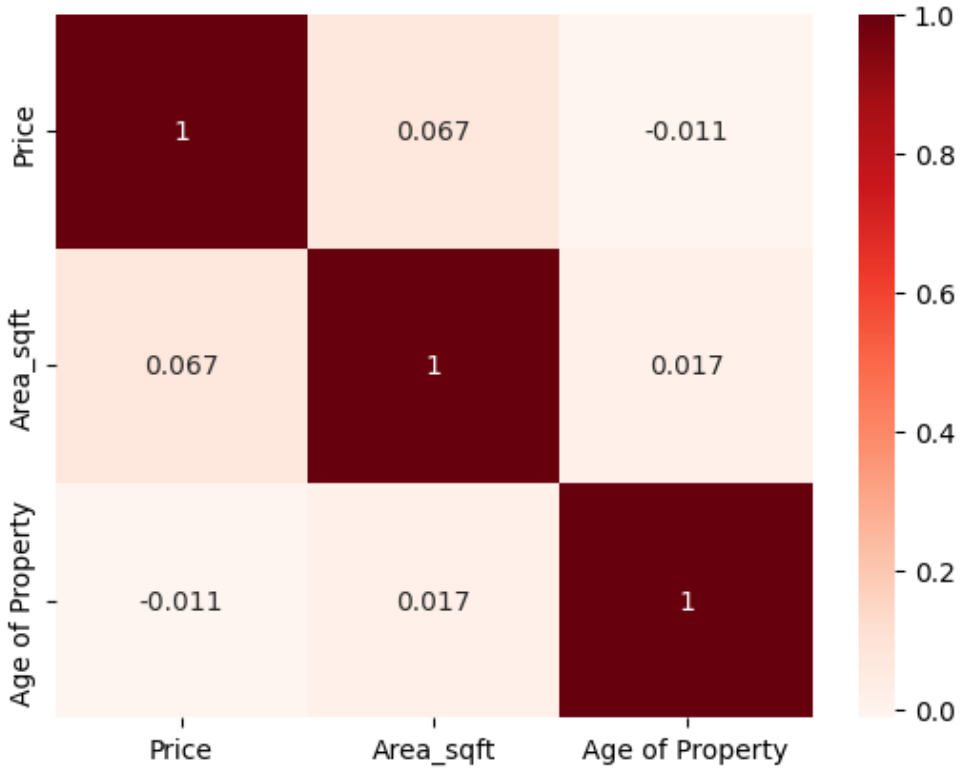

### Barplot of Count of Listing by Area_sqft



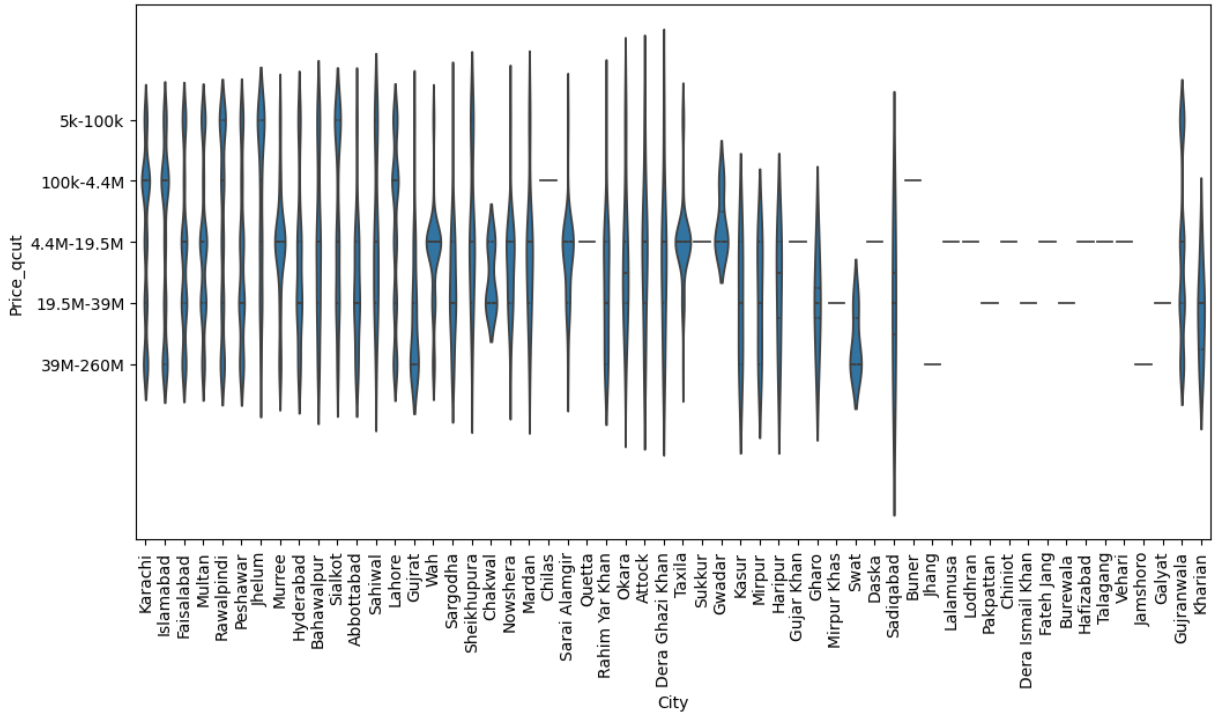
**Count plot of Age of Property**



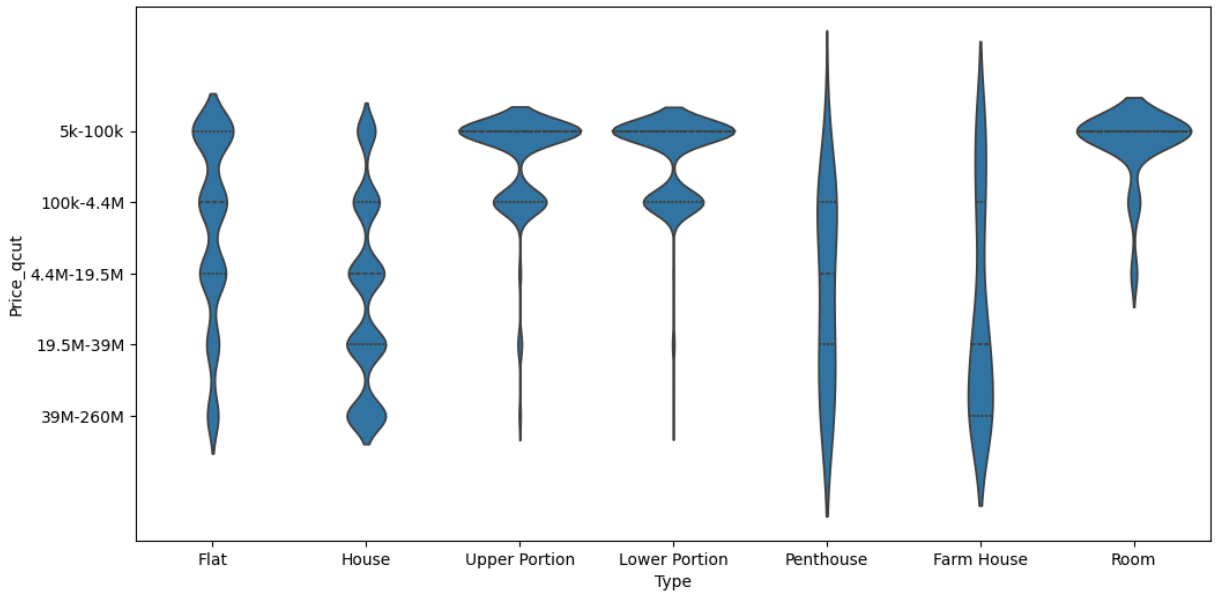

**Correlation Heatmap of Price, Area_sqft and Age of Property**



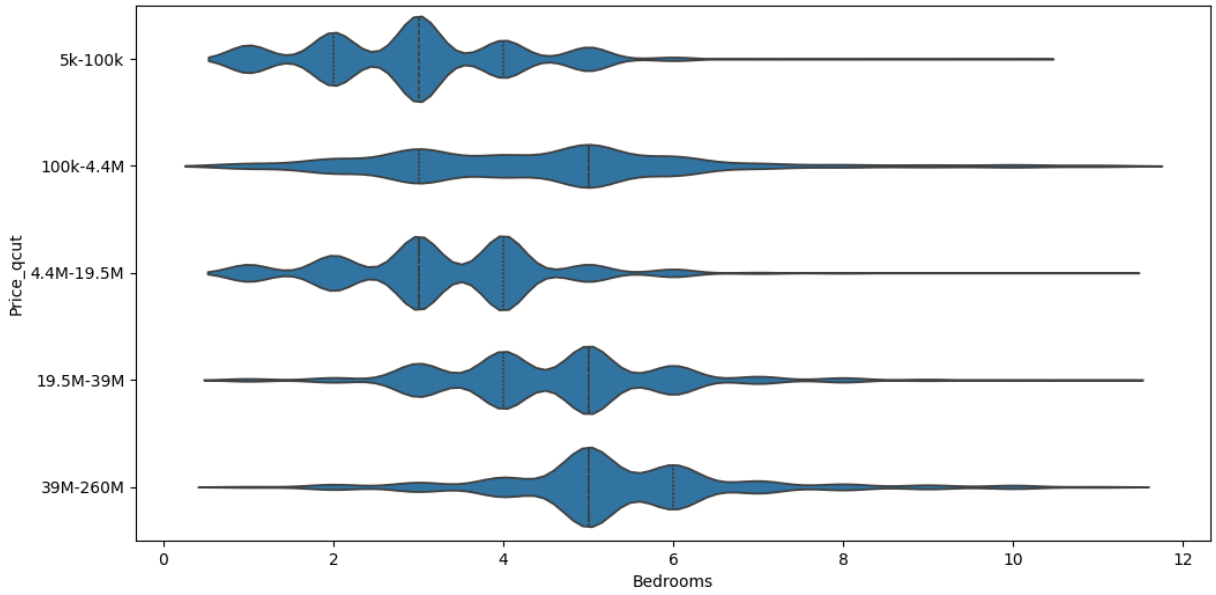
## Violin plot of Price by City




## Violin Plot of Price by Property Type



**Violin plot of Price by Bedrooms**




## Section 4: Data Interpretation

### Insights and Recommendations

1. The highest number of property listings (around 3800) are listed in the price range 5k-100k which are for rent. All the listings from 3000-3700 are listed between the price range 5k-260M which includes rental and for sale properties. The number of listings are equally distributed in all price ranges. (barplot of count of listings by price)
2. Around 5100 property listings have an area of 71-1360 sqft. The lowest quantity of around 1900 listings have an area of 1360-1998 sqft. Around 2600 listings have the highest area of over 5445 sqft. (barplot of count of listings by area)
3. Most of the listings on zameen are aged 0-3 years from current year 2025. It indicates that most of the projects were completed in recent years and now they are up for rent or sale. (barplot of age of property)
4. Area sqft and Age of Property show a weak correlation with Price and between each other. Price is not increasing area size and age of property. (heatmap)
5. The violin plot for property prices in various cities shows cities Islamabad, Karachi, Lahore, Gujranwala, Sialkot, Faisalabad and Rawalpindi have the highest real estate rates for rent and sale. The long vertical line for city Islamabad indicates that price variation is the greatest in Islamabad. Given the higher population and status as the main economic centres of the country, property demand is greater and the business is booming there. (violin plot prices and cities)

6. In smaller cities like Chilas, Buner, Jhang, etc. the flat horizontal line indicates lesser price variation and that property rates are in the middle of the total price range in this dataset. Property is cheaper because the demand for developing and purchasing property is lower in small undeveloped cities including the northern areas. (violin plot prices and cities)
7. Room listings are mostly concentrated in the price range 5k-100k which is primarily for rent. The farmhouses and penthouses are thinly stretched between low and high price range. It indicates that fewer listings are available for rent and sale. Houses and flats are equally spread in all price ranges which indicates that most people are in the market for houses and flats for sale and rent. Upper and lower portion listings are in the market for rent primarily which is exhibited by the price range between 5k-1M. (violin plot for prices and property type)
8. For 5k-100k, 2-4 bedrooms are listed for rent and between 4.4M-19.5M they are listed for sale. 4-6 bedroom listings are priced between 19.5M-260M for sale. This data also indicates that 1-6 bedrooms are most common in listings which indicates demand for it. A large portion of 4 bedroom listings are also priced between 19.5-39M. (violin plot of prices and number of bedrooms)

**Summary of Learnings**

1. Property prices are most definitely determined by location, size and type of property. Big cities have more expensive properties than small cities
2. However, prices are not correlated to area and age of property as per the data
3. Most of the prices on zameen are also based on speculation. The sellers distort the actual property prices to increase their profits and property prices in future based on mere guesswork
4. In the absence of a clear price framework, we cannot clean the inflated property prices which is a clear challenge in analysing the dataset

**Suggestions for Stakeholders (Investors)**

1. Stakeholders should make effort towards creating a regulated real estate market to prevent speculative pricing
2. Invest in major cities of Pakistan where property rates are higher due to better administration and access to amenities

3. Despite, the slow market, there is a demand for property but pricing should be reasonable to improve sales
4. Many previous projects that were completed in the past 3 years will struggle to sell because of high cost of construction and low purchasing power in the market. Hence, consider all the costs and development time before starting a new project or investing in developing projects